CROP: Contextual Region-Oriented Visual Token Pruning

Jiawei Guo[1,2], Feifei Zhai[✉1], Pu Jian[1,2], Qianrun Wei[1,2], Yu Zhou[1,3]
[1] State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, CAS, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Fanyu AI Laboratory, Zhongke Fanyu Technology Co., Ltd, Beijing, China
{guojiawei2024, feifei.zhai, jianpu2023, weiqianrun2025}@ia.ac.cn, yzhou@nlpr.ia.ac.cn

Abstract

Current VLM-based VQA methods often process entire images, leading to excessive visual tokens that include redundant information irrelevant to the posed question. This abundance of unnecessary image details creates numerous visual tokens, drastically increasing memory and computational requirements in VLMs. To address this, we propose Contextual Region-Oriented Visual Token Pruning (CROP[1]), a novel framework to compress visual tokens through a two-step process: Localization and Pruning. Specifically, CROP first employs an efficient model to identify the contextual region relevant to the input query. Subsequently, two distinct strategies are introduced for pruning: (1) Pre-LLM Compression (PLC), which adaptively compresses different image regions with varying ratios, and (2) Inner-LLM Pruning (ILP), a training-free method that prunes tokens within early LLM layers guided by the identified contextual region. Extensive experiments on a wide range of VQA tasks demonstrate that CROP significantly outperforms existing visual token pruning methods and achieves state-of-the-art performance.

1 Introduction

Recently, visual question answering (VQA) have achieved remarkable progress due to the rapid development of vision language Models (VLMs) (Bai et al., 2023; Yin et al., 2023; Li et al., 2024b; Ren et al., 2024; Singh et al., 2019; Zhou et al., 2025). Current VLM-based VQA methods utilize information from the entire image, but for specific questions, we need to locate local image regions to support the answer. Moreover, redundant image information also introduce a large number of visual tokens, requiring much higher memory and computation in VLMs (Zhang et al., 2025b; Huang et al., 2024). For example, there are 576 visual tokens in LLaVA-1.5 (Liu et al., 2023), and the number is 2880 for a 672*672 image in LLaVA-NeXT (Yang et al., 2024).

For example in Figure 1(a), we actually only need a small region to answer the question, but still have to transform the entire image into so many visual tokens. To overcome this problem, visual token pruning methods have been emerged. Some methods reduce visual tokens before inputting them into LLM (Huang et al., 2024), which primarily depend on intrinsic image semantics. Others perform pruning inside the early LLM layers, usually on the basis of attention map (Chen et al., 2024). However, these methods fail to account for the input question, which might ignore the key task-relevant information (Li et al., 2024a).

In this paper, we introduce a novel framework, Contextual Region-Oriented Vision Token Pruning (CROP), to facilitate effective visual token pruning. We define the contextual region as a contiguous visual area of the input image that captures the key information for answering the question. The proposed CROP framework comprises two stages: Localization and Pruning. During the Localization Stage, an efficient model identifies the contextual region. Subsequently, in the Pruning Stage, this identified region serves as a key information source to guide two simple yet effective pruning strategies: Pre-LLM Compression (PLC) and Inner-LLM Pruning (ILP).

In PLC method, we propose a compression module that adaptively adjusts the compression ratio, assigning a lower compression ratio to the contextual region obtained from the localization stage, and a higher ratio otherwise. The ILP method prunes visual tokens in the LLM layers by the guidance of L-CR. Extensive experiments on a wide range of VQA benchmarks show that the proposed PLC and ILP methods consistently outperform the existing compression techniques, and achieve the state-of-

✉Corresponding Author
[1]The related code and dataset are released at: https://github.com/JiaweiGuo98/CROP

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 9756–9772
November 4-9, 2025

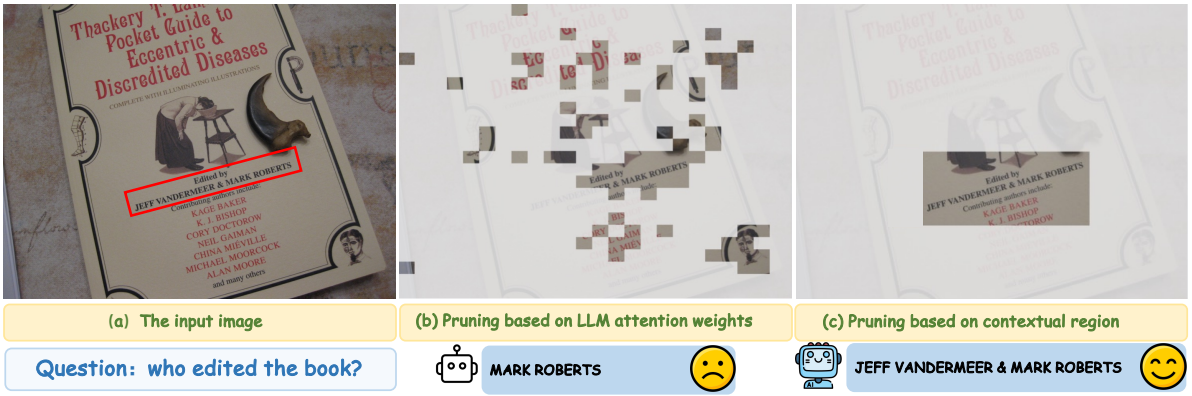

Figure 1: Comparison of different visual token pruning. (a) The input image and question. The red rectangle shows the region containing the answer of the question. (b) The visualization of retained tokens by attention-based pruning method. (c) The visualization of identified contextual region by our CROP framework.

the-art performance.

The main contributions of this paper are summarized as follows:

- We develop an efficient Localization model to identify contextual regions, which serve as a plug-and-play component for guiding visual token pruning.
- We introduce a novel contextual region-oriented vision token pruning framework, and develop two different approaches on pruning, PLC and ILP. To our best knowledge, this is the first work of introducing contextual regions for visual tokens pruning.
- Experimental results demonstrate that our proposed method achieves state-of-the-art performance without requiring any training or fine-tuning in ILP method.

2 Related Works

2.1 Large Language Models

The advancement of Large Language Models (LLMs) has redefined state-of-the-art performance across a vast landscape of tasks, spanning foundational natural language understanding (Jing and Zhao, 2024; Zhang et al., 2025c; Chen et al., 2025b,a), machine translation (Guan et al., 2025; Liang et al., 2024; Zhang et al., 2025d), and complex reasoning in domains like mathematics and agent-based systems (Sun et al., 2025; Xu et al., 2025). A particularly impactful frontier has been their expansion into multimodality, enabling breakthroughs in audio processing (Diao et al., 2025a, 2024, 2025b) and, crucially for this work, visual understanding (Jian et al., 2025a,b, 2024), which has given rise to powerful VLMs. While these VLMs excel at many tasks, their success comes at a high computational cost due to the large number of visual tokens processed from each image, motivating the need for efficient token reduction strategies.

2.2 Vision Token Reduction in VLMs

In recent years, a substantial amount of research has emerged focusing on visual token pruning and compression, with these methods emphasizing the use of attention mechanisms within VLMs to retain important visual information (Han et al., 2025; Ye et al., 2025; Yan et al., 2024; Xing et al., 2024; Liu et al., 2024b). For instance, FastV (Chen et al., 2024) leverages cross-modal attention from intermediate layers of the model to preserve the Top-R visual tokens, while TwigVLM (Shao et al., 2025) introduces additional trainable layers to improve the precision of token selection. VisionZip (Yang et al., 2024) aims to use visual encoder attention to retain, compress, and merge key visual information. SparseVLM (Zhang et al., 2024) employs bidirectional selection of both text and visual tokens to extract more informative visual representations. LLaVA-Mini (Zhang et al., 2025b) incorporates an additional fusion module between the encoder and the LLM to perform early cross-modal fusion. However, these methods often result in discrete and fragmented visual tokens, disrupting the spatial semantics and continuity of visual regions. In contrast, our CROP method preserves continuous contextual regions directly relevant to the input question

2.3 Contextual Region Perception in VLMs

Recent advancements have shown that VLMs possess strong capabilities in identifying and grounding information within specific visual regions. For

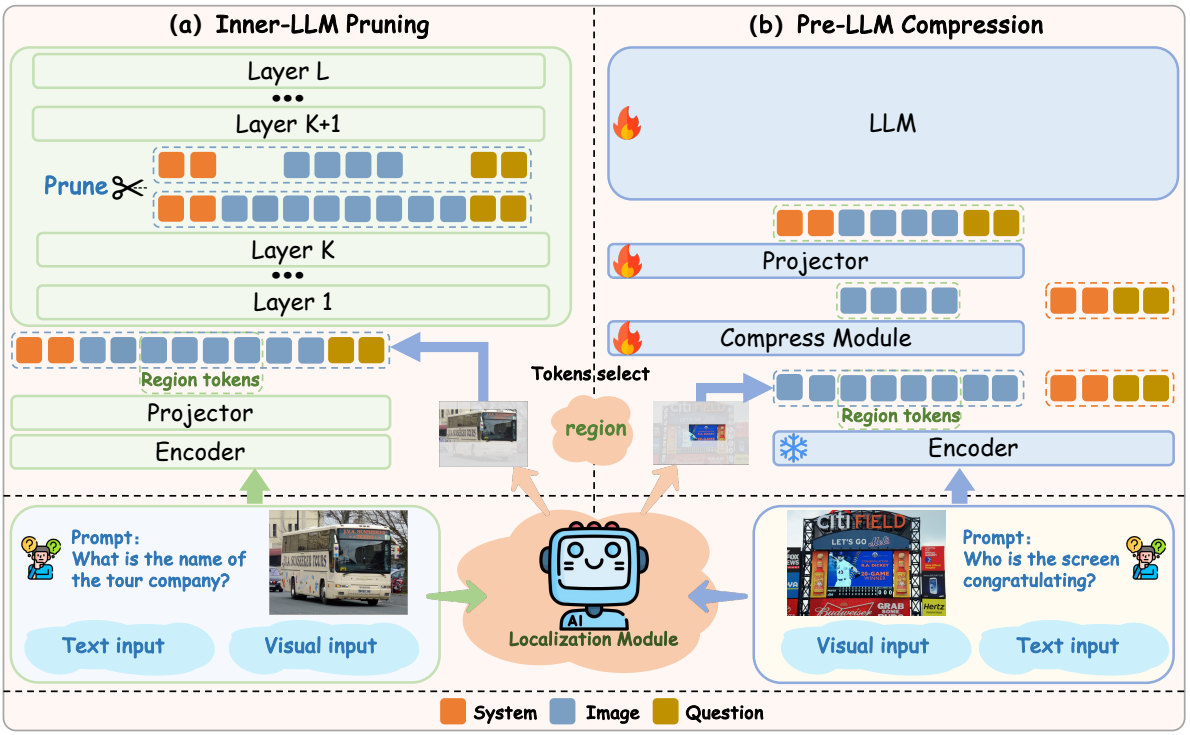

Figure 2: Overview of the CROP framework. Given achen2025lr2benchevaluatinglongchainreflective visual input and a textual prompt, the Localization Module first identifies task-relevant contextual regions. This regional information then guides two distinct visual token compression strategies: (a) ILP, where tokens outside the contextual region are pruned directly within an early layer (Layer K) of the main VLM without retraining it; and (b) PLC, where a dedicated compression module processes tokens from contextual and non-contextual regions before they are fed to the LLM. The visual input pipeline and the LLM are shown in context for both strategies.

example, Zhang et al. (2025a) demonstrated that VLMs often know which visual areas to focus on, even when answering questions incorrectly. Shao et al. (2024) found that cropping and re-inputting relevant regions enhances visual perception, introducing datasets for localization. Additionally, VPT (Yu et al., 2025) replaced precise coordinates with region selection tokens, improving localization accuracy. These studies suggest that VLMs have significant abilities in question-guided region localization (Li et al., 2025). Our work builds on this by fine-tuning VLMs for explicit region identification, guiding visual token pruning and compression to preserve the most contextually relevant visual information.

2.4 Chain-of-thought Reasoning

Chain-of-Thought (Wei et al., 2023) (CoT) prompting has demonstrated that intermediate reasoning steps improve problem-solving in large language models. While this approach has been extended to the vision domain, VLMs still show limited ability to effectively localize and interpret visual regions. To overcome this limitation, recent methods have introduced locate-then-answer paradigms. For example, Wu and Xie (2023) fine-tunes VLMs to use a visual search model when additional localization is necessary, and Luan et al. (2024) mimics human scanning by first generating a global description before localizing regions. Other approaches, such as Man et al. (2025); Ge et al. (2025), employ Mixture-of-Experts image encoder or cross-attention across VLMs to refine predictions. However, these methods often rely on discrete attention mechanisms or precise bounding box regression, which can struggle with sustained focus and require costly model-specific fine-tuning. In contrast, our approach uses a lightweight VLM for localization that predicts coordinates at a coarse, block-level granularity. This strategy not only achieves more robust localization but also allows our module to be integrated into various VLMs without requiring any retraining.

3 Method

3.1 Overview

To preserve the completeness of contextual regional information during visual token processing—a factor we hypothesize is crucial for robust understanding compared to methods that might select sparse,

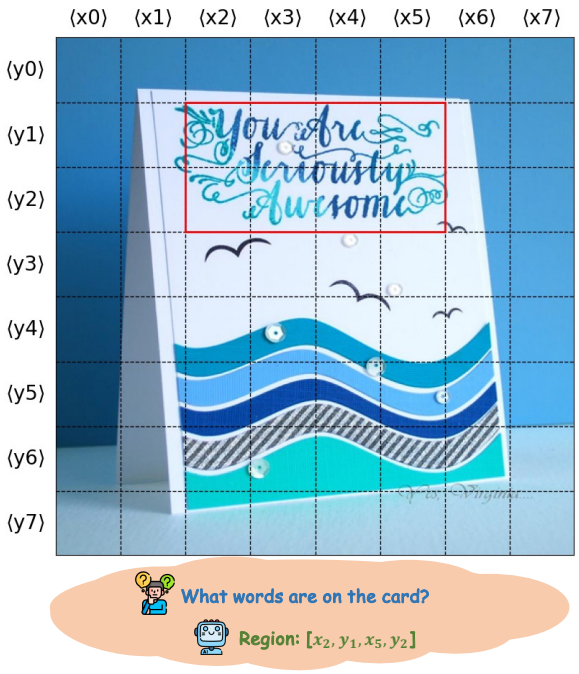

Figure 3: Illustration of the 8×8 Grid-based Region Definition for Contextual Region Localization. The input image is overlaid with an 8×8 grid. For a given question (e.g., "What words are on the card?"), the Localization Module identifies the contextual region containing the answer (e.g., the text "You Are Seriously Awesome"). This region is represented by the minimum and maximum 0-indexed block coordinates $[x_{min}, y_{min}, x_{max}, y_{max}]$, which in this example corresponds to $[x_2, y_1, x_5, y_2]$.

disconnected tokens—we introduce CROP. The overall architecture of CROP is illustrated in Figure 2. The framework operates in two main stages: first, a dedicated Localization Module identifies contextual regions pertinent to a given query. Second, this regional information guides the subsequent visual token processing in the main VLMs. This approach emphasizes retaining comprehensive visual information within these contextual regions, particularly their inherent spatial structures and the relationships between constituent patches.

The main VLMs used in our experiments typically process an input image into a grid of visual tokens (e.g., 24×24=576 tokens from ViT-L/14 (Radford et al., 2021)), which maintain spatial correspondence with the original image patches. This spatial mapping is crucial as it allows CROP to accurately translate identified contextual regions into specific sets of visual tokens for targeted processing. Losing this explicit mapping, as might occur in some global pooling or aggressive token merging strategies, could obscure fine-grained spatial details vital for many VQA tasks. We instantiate and evaluate CROP through two distinct strategies: ILP and PLC. ILP directly removes visual tokens from non-contextual regions within the early layers of the main VLMs, requiring no retraining. PLC employs a lightweight module to compress visual tokens before they enter the LLM, prioritizing the fidelity of contextual regional information while integrating broader context.

3.2 Contextual Region Localization

A lightweight VLM serves as an independent Localization Module, which pre-identifies contextual regions to provide guidance to the main VLM. Direct bounding box generation by VLMs can be challenging, especially across varied image resolutions. However, guiding models to identify broader regional extents can yield more robust localization. Inspired by this and approaches like Visual Prompt Tuning (Yu et al., 2025) and Visual-cot (Shao et al., 2024), we developed a specialized training dataset to enhance the Localization Module's ability to discern salient information based on spatial layout.

We represent images on an 8×8 grid, as depicted in Figure 3. A rectangular contextual region within this grid is defined by block coordinates $[x_{min}, y_{min}, x_{max}, y_{max}]$. These are 0-indexed, meaning each coordinate can range from 0 to 7 inclusive, and must satisfy $0 \leq x_{min} \leq x_{max} \leq 7$ and $0 \leq y_{min} \leq y_{max} \leq 7$.

We selected the Qwen2VL-2B model (Wang et al., 2024) as our Localization Module and fine-tuned it on this curated dataset. This module, requiring approximately 8 hours of training, achieves high localization accuracy and can be seamlessly integrated as a plug-and-play component with various large VLMs.

3.3 Visual Token Compression

Given the localized contextual regions, CROP implements ILP and PLC. Our experiments primarily utilize the LLaVA model family (Liu et al., 2023, 2024a), which employs a ViT (Dosovitskiy et al., 2021) for visual encoding. In such models, an input image I^v is processed by the visual encoder, and its output features are transformed by a projector (e.g., an MLP) into N^2 visual tokens, $X^v \in \mathbb{R}^{N^2 \times d_h}$. These visual tokens are projected into a space compatible with the text prompt tokens X^T, where d_h is the LLM's hidden embedding dimension. The combined multimodal input for the LLM often follows the structure:

$$\langle X^q_1, \ldots, X^q_m, X^v_1, \ldots, X^v_{N^2}, X^q_{m+1}, \ldots, X^q_{N_t} \rangle \quad (1)$$

Function	Dataset
Region Localization (410k)	TextVQA (73k), Flickr30k (136k), DocVQA (33k) VSR (3k), GQA (88k), OpenImage (43k) CUB (4k), V7W (30k)
Preserving Instruction Following Ability (200k)	LLaVA Instruction Tuning data(200k)

Table 1: Composition of the training dataset. Our training dataset includes the Region Localization samples used to train the Localization and Compress Modules, as well as the LLaVA Instruction Tuning data aimed at preserving the model's original instruction-following capability. In total, the training dataset comprises 610k samples.

where $\langle X_1^q, \ldots, X_m^q \rangle$ is the system prompt, and $\langle X_{m+1}^q, \ldots, X_{N_t}^q \rangle$ is the user query.

Inner-LLM Pruning. While some existing methods for reducing visual tokens in VLMs utilize cross-modal attention or introduce additional trainable modules for token selection, our ILP offers a distinct approach. Our observations, consistent with existing literature (Chen et al., 2024, 2023), suggest that VLMs rely more heavily on explicit visual tokens in their earlier layers, with information becoming progressively more abstract and multimodally fused in deeper layers.

ILP leverages this by using the precise, task-relevant region information from the Localization Module. Based on the identified contextual regions, visual tokens falling outside this area are removed at a designated early layer K of the VLM, as illustrated in Figure 2(a). This approach is plug-and-play, as it requires no retraining of the large VLM. Our experiments validate that this early-stage, region-guided pruning achieves state-of-the-art performance on multiple VQA benchmarks, even on VLMs not specifically adapted for CROP.

Pre-LLM Compression. We hypothesize that contextual regions contain the most critical visual information for query resolution. To explore this while retaining some global context, PLC is designed as an external, trainable compression module that prioritizes contextual region fidelity, depicted in Figure 2(b).

Visual tokens from the encoder, X^v, are first partitioned into contextual region tokens X^{kv} and non-contextual region tokens X^{nkv}, guided by the Localization Module. We introduce two sets of learnable queries, Q^k and Q^{nk}, to compress these token sets respectively. Positional encodings $P(\cdot)$ are added to queries and tokens before attention. Compression is achieved via scaled dot-product attention (Equations 2 and 3):

$$\hat{X}^{kv} = \text{softmax}\left(\frac{P(Q^k)(P(X^{kv}))^T}{\sqrt{d_h}}\right) P(X^{kv}) \tag{2}$$

$$\hat{X}^{nkv} = \text{softmax}\left(\frac{P(Q^{nk})(P(X^{nkv}))^T}{\sqrt{d_h}}\right) P(X^{nkv}) \tag{3}$$

Here, $\hat{X}^{kv}$ and $\hat{X}^{nkv}$ are the compressed representations.

In addition to the compressed representations, we define anchor tokens X^r to preserve fine-grained details. X^r is an 8x8 square patch of 64 tokens, extracted from the original visual encoder output, centered around the geometric midpoint of the contextual region X^{kv}. These query the compressed contextual region tokens $\hat{X}^{kv}$ for contextual integration, followed by a residual connection:

$$X^{fused} = \text{softmax}\left(\frac{P(X^r)(P(\hat{X}^{kv}))^T}{\sqrt{d_h}}\right) P(\hat{X}^{kv}) + X^r \tag{4}$$

The final compressed visual representation $\hat{X}^v$ concatenates the fused contextual region tokens with the compressed non-contextual region tokens:

$$\hat{X}^v = \text{Concat}(X^{fused}, \hat{X}^{nkv}) \tag{5}$$

This PLC strategy aims to balance substantial token reduction with the preservation of essential regional details and broader contextual cues.

4 Experiments

4.1 Training Details

We constructed the training dataset based on the datasets from LLaVA (Liu et al., 2023) and Visual-cot (Shao et al., 2024). The composition of this dataset is detailed in Table 1. All training procedures for CROP were conducted on a server equipped with four NVIDIA GeForce RTX A100 GPUs. The training process consists of two main stages:

Method	LLaVA-1.5-7B							LLaVA-1.5-13B						
	GQA	MME	VQA^{T}	SQA	VQA^{V2}	POPE	Avg.	GQA	MME	VQA^{T}	SQA	VQA^{V2}	POPE	Avg.
Upper Bound, 576 Tokens (**100%**)														
LLaVA-1.5	61.9	1862	58.2	69.5	78.5	85.9	100%	63.2	1818	61.3	72.8	80.0	85.9	100%
Retain Averaged 192 Tokens (↓ **66.7%**)														
FastV	56.5	1786	57.3	**69.5**	74.6	79.2	95.5%	60.3	1807	60.4	74.0	77.7	82.3	98.0%
SparseVLM	57.6	1721	56.1	69.1	75.6	83.6	95.8%	58.7	1768	45.4	73.1	-	82.2	92.1%
PDrop	57.3	1797	56.5	69.2	75.1	82.3	96.2%	61.3	1663	**60.7**	73.6	78.7	84.8	97.6%
VisionZip	59.3	1783	57.3	68.9	76.8	**85.3**	97.7%	59.1	1754	59.5	73.5	78.1	85.1	97.5%
VisionZip‡	60.1	**1834**	**57.8**	68.2	77.4	84.9	98.4%	61.6	1790	59.9	72.7	78.6	84.5	98.4%
CROP-ILP	**61.3**	1817	57.4	69.2	**77.7**	**85.3**	**98.9%**	**62.7**	**1822**	60.4	**74.2**	**78.9**	**86.2**	**99.8%**
Retain Averaged 128 Tokens (↓ **77.8%**)														
FastV	53.0	1646	56.0	**69.5**	69.2	73.2	90.6%	57.5	1758	58	73.8	74.3	79.3	94.8%
SparseVLM	56.0	1696	54.9	67.1	73.8	80.5	93.4%	57.9	**1774**	49.9	69.9	-	81.1	92.2%
PDrop	57.1	1761	56.6	68.4	72.9	82.3	95.2%	61.0	1490	**60.2**	73.3	**78.2**	83.6	95.4%
VisionZip	57.6	1762	56.8	68.9	75.6	83.2	96.3%	57.9	1743	58.7	74.0	76.8	85.2	96.7%
VisionZip‡	58.9	**1823**	**57.0**	68.3	76.6	83.7	97.4%	60.1	1736	59.2	73.0	77.6	83.8	97.0%
CROP-ILP	**60.8**	1771	56.9	**69.5**	**76.8**	**84.4**	**97.9%**	**61.6**	1768	59.4	**74.1**	78.0	**85.8**	**98.5%**
Retain Averaged 64 Tokens (↓ **88.9%**)														
FastV	44.1	1218	50.7	70.0	52.0	55.6	75.9%	50.1	1408	52.2	73.2	61.1	69.3	83.3%
SparseVLM	52.7	1505	51.8	62.2	68.2	75.1	86.5%	50.6	1402	22.7	69.0	-	65.0	72.9%
PDrop	47.5	1561	50.6	69.0	69.2	55.9	83.3%	54.1	1247	55.3	73.1	70.8	66.1	85.0%
VisionZip	55.1	1690	55.5	69.0	72.4	77.0	92.7%	56.2	1676	57.4	**74.4**	73.7	76.0	92.9%
VisionZip‡	57.0	**1756**	**56.0**	68.8	74.2	80.9	95.1%	58.1	1671	**58.5**	72.3	75.2	81.6	94.6%
CROP-ILP	59.6	1675	54.9	71.5	74.8	**83.6**	**96.0%**	60.4	**1708**	56.8	73.8	**76.0**	**84.8**	**96.2%**
CROP-PLC	**60.3**	1634	55.2	**71.8**	**75.0**	80.2	95.4%	**61.1**	1693	57.2	73.9	75.7	81.8	95.8%

Table 2: **Performance of CROP on LLaVA-1.5 compared to existing methods** under three different pruning ratios. The **bold numbers** indicate the best performance achieved by each MLLM, and the underline numbers are the second best. Specifically, VisionZip‡refers to fine-tuned version of VisionZip, where the projector has been fine-tuned to align the pruned visual tokens with the semantic space of the LLM.

Stage 1: Localization Module Training The Qwen2VL-2B model was fine-tuned to function as the Localization Module. During this stage, the visual encoder of Qwen2VL-2B was kept frozen, while all other components were trained on our Region Localization data and Preserving Instruction Following Ability data. Fine-tuning was performed for 2 epochs using the AdamW optimizer with a learning rate of 2e-5. This stage typically completed in approximately 12 hours.

Stage 2: Pre-LLM Compression Module and VLM Co-Fine-tuning For the PLC strategy, the LLaVA-1.5 model was co-fine-tuned with the newly introduced PLC components. As in Stage 1, LLaVA's visual encoder remained frozen. The learnable parameters of the PLC module, along with LLaVA's projector and LLM backbone, were jointly optimized. This co-fine-tuning utilized our complete training set. Specifically, Region Localization samples with ground truth bounding box annotations were used to train the compression module and help the LLM adapt to the compressed visual inputs, while the LLaVA Instruction Tuning data was used to preserve the model's original instruction-following capability. Fine-tuning was conducted for 2 epochs using the AdamW optimizer with a learning rate of 1e-5. This stage typically required approximately 26 hours.

4.2 Experimental Setup

We evaluated our CROP method on LLaVA-1.5-7B and LLaVA-1.5-13B, conducting extensive ablation studies to verify the importance of contextual region preservation for model perception. Experiments were performed across multiple benchmarks (see Appendix for dataset and metric details), with all evaluations adhering to official dataset guidelines and LLaVA's metrics.

For the ILP strategy, pruning was implemented at layer K=2 of the target VLM. Based on the critical regions identified by our Localization Module, visual token pruning rates were averaged to 66.7%, 77.8%, and 88.9% across the test set. A crucial aspect of this setup is that the retained tokens consis-

tently form continuous rectangular image regions.

For the PLC strategy, the number of learnable queries for contextual regions N_{Q^k} and non-contextual regions $N_{Q^{nk}}$ were configured to 64 and 4, respectively. The number of anchor-preserved tokens N_r from the contextual region was set to 64.

4.3 Main Results

Results on LLaVA-1.5. The primary performance metrics of CROP are presented in Table 2. Across the tested pruning ratios, CROP consistently maintained robust visual comprehension capabilities, achieving SOTA results on several benchmarks.

With the ILP strategy, at a 66.7% visual token pruning rate, LLaVA-1.5-7B and LLaVA-1.5-13B retained 98.9% and 99.8% of their respective baseline performances. Even under a more aggressive 88.9% pruning rate, these models maintained 96.0% and 96.2% of their original efficacy. Notably, these ILP results were achieved without any fine-tuning of the LLaVA models.

Our PLC strategy, evaluated on LLaVA-1.5, demonstrated 95.4% and 95.8% performance retention relative to its baseline at an 88.9% visual token pruning rate. These findings strongly validate the effectiveness of CROP, underscoring the benefits of preserving contextual regional information in visual token compression.

Results on LLaVA-NeXT. To further validate the effectiveness of the proposed CROP framework, we conduct experiments on the LLaVA-NeXT series of models. Unlike LLaVA-1.5, LLaVA-NeXT partitions each input image into four sub-images, which are then processed alongside the original image by the visual encoder—effectively increasing the input to five images per instance. While this design enhances visual perception, it also introduces substantial redundancy and computational overhead. To improve inference efficiency, we apply contextual region-based pruning to the visual tokens encoded from both the original image and its sub-images. We evaluate CROP under three different token retention settings to systematically assess its performance advantages. As shown in Table 3, our proposed CROP framework consistently maintains strong performances.

Method	GQA	MME	VQA^{T}	SQA	VQA^{V2}	RelAcc
Upper Bound, 576 Tokens (**100%**)						
LLaVA-NeXT	64.2	1842	61.3	70.2	80.1	100%
Retain Averaged 640 Tokens (↓ **77.8%**)						
FastV	62.0	**1807**	60	69.1	79.5	98.0%
SparseVLM	60.3	1772	57.8	67.7	77.1	95.4%
VisionZip	61.3	1787	60.2	68.1	79.1	97.3%
VisionZip‡	62.4	1778	**60.8**	67.9	79.9	97.9%
CROP-ILP	**63.2**	1768	60.0	**69.6**	**80.0**	**98.3%**
Retain Averaged 320 Tokens (↓ **88.9%**)						
FastV	54.9	1539	54.8	68.2	69.6	88.5%
SparseVLM	57.7	1694	55.9	67.3	73.4	92.1%
VisionZip	59.3	1702	58.9	67.3	76.2	94.4%
VisionZip‡	61.0	**1770**	**59.3**	67.5	78.4	96.4%
CROP-ILP	**62.0**	1749	59.1	**69.1**	**78.7**	**96.9%**
Retain Averaged 160 Tokens (↓ **94.4%**)						
FastV	49.3	1496	47.5	67.9	68.2	83.5%
SparseVLM	51.2	1542	46.4	67.5	66.3	83.6%
VisionZip	55.5	1630	56.2	68.3	71.4	90.6%
VisionZip‡	58.2	**1699**	57.3	67.5	75.6	93.4%
CROP-ILP	**60.2**	1668	**57.7**	**68.7**	**76.1**	**94.3%**

Table 3: Performance of CROP-ILP with LLaVA-NeXT-7B on Various VLM Benchmarks. The **bold values** indicate the best performance.

4.4 Efficiency Analysis

Owing to the workflow design of CROP, the Localization Module and the VLM backbone operate relatively independently during inference. While the Localization Module introduces a non-negligible number of parameters, it remains lightweight compared to the VLM backbone and achieves faster inference. Consequently, when handling multiple batches or video streams, the VLM backbone only needs to wait for the Localization Module once, typically during the first batch. After this initial step, the localization module and the VLM backbone can process subsequent inputs in parallel. Thanks to this "ahead-of-time prediction" property, the VLM backbone is freed from serial waiting for localization outputs, thereby reducing overall latency. Following the efficiency analysis setup of prior work, we mainly measured image encoding time, KV cache load time, and transformers forward time, while excluding model loading time and other overheads. To validate this claim, we conducted speed benchmarking experiments on the LLaVA-NeXT model. As shown in the Table 4, our ILP method accelerates LLaVA-NeXT-13B by more than twofold and even surpasses the speed of the smaller LLaVA-NeXT-7B. These results demonstrate the effectiveness of CROP in enhancing efficiency.

5 Analyses and Discussion

Previous visual token compression methods have largely overlooked the importance of preserving

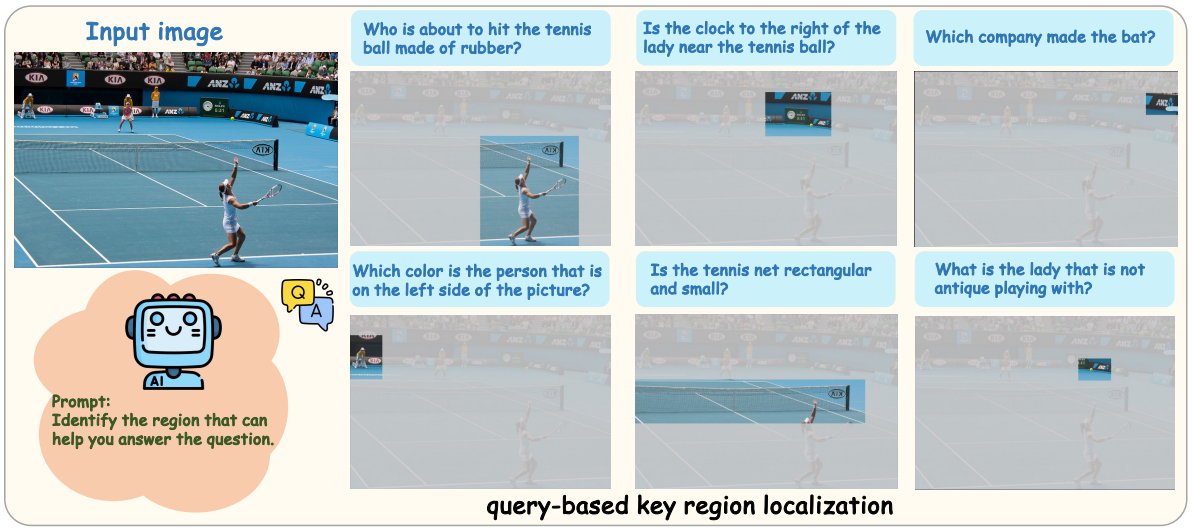

Figure 4: Qualitative examples of query-based contextual region localization by our Localization Module. For different questions, the module identifies specific, contiguous regions (highlighted) crucial for answering, demonstrating focused attention compared to potentially diffuse attention in other methods.

Method	GQA	MME	VQA^{T}	SQA	VQA^{V2}
Upper Bound, 576 Tokens **(100%)**					
LLaVA-NeXT-7B	3356s	606s	1611s	874s	2336s
LLaVA-NeXT-13B	5394s	953s	2497s	1349s	3671s
Retain Averaged 640 Tokens (↓ 77.8%)					
FastV	2857s	516s	1401s	860s	1928s
CROP-ILP	2833s	497s	1400s	850s	1861s
Retain Averaged 160 Tokens (↓ 94.4%)					
FastV	1973s	350s	1029s	580s	1303s
CROP-ILP	2028s	355s	1066s	592s	1330s

Table 4: Efficiency analysis of CROP-ILP on LLaVA-NeXT-13B. The table reports the total inference time on multiple VQA benchmarks, comparing the baseline LLaVA-NeXT-7B/13B with FastV and our CROP-ILP applied to LLaVA-NeXT-13B.

Method	GQA	MME	VQA^{T}	SQA	POPE	RelAcc
Upper Bound, 576 Tokens **(100%)**						
LLaVA-1.5	61.9	1862	58.2	69.5	85.9	100%
Retain Averaged 192 Tokens (↓ 66.7%)						
prune	58.7	1735	56.8	**69.2**	**85.4**	96.9%
CROP-ILP	**61.3**	**1817**	**57.4**	**69.2**	85.3	**98.8%**
Retain Averaged 128 Tokens (↓ 77.8%)						
prune	57.9	1734	56.2	**69.6**	**84.6**	96.4%
CROP-ILP	**60.8**	**1771**	**56.9**	69.5	84.4	**97.9%**
Retain Averaged 64 Tokens (↓ 88.9%)						
prune	55.0	1625	54.1	69.9	**83.8**	93.4%
CROP-ILP	**59.6**	**1675**	**54.9**	**71.5**	83.6	**96.2%**

Table 5: Performance of LLaVA-1.5-7B with "prune" Guided by the Localization Module. Results are shown across various VQA benchmarks and pruning rates, compared to baseline LLaVA-1.5-7B and representative discrete token pruning methods. The **bold values** indicate the best performance.

contextual regions during compression. This section analyzes the impact of our CROP strategy on model perception and understanding, and further evaluates the effectiveness of our Localization Module.

5.1 Localization Module's Token Selection

To validate the efficacy of our Localization Module, we conducted experiments on LLaVA-1.5-7B. Using the contextual regions identified by the module, we derived the corresponding visual token indices. Tokens outside these contextual regions were pruned entirely before being fed to the LLM, an approach we denote as **prune**. We tested this at visual token pruning rates of 66.7%, 77.8%, and 88.9%.

As shown in Table 5, even when relying solely on visual information from these localized contextual regions, the model maintained robust performance. Notably, this simple region-only approach surpassed several meticulously designed discrete token pruning strategies. These results indicate two key findings: (1) our Localization Module accurately identifies critical, query-relevant regions across diverse benchmarks, and (2) retaining tokens from these contextual regions preserves the majority of essential visual information more effectively than discrete token selection methods.

5.2 Visual Token Selection Strategies

Discrete token pruning strategies, especially those reliant on cross-modal attention, can suffer from attention diffusion. This may lead to the retention of

Method	GQA	MME	VQA^T	SQA	POPE	RelAcc
Upper Bound, 576 Tokens **(100%)**						
LLaVA-1.5	61.9	1862	58.2	69.5	85.9	100%
Retain Averaged 64 Tokens ($\downarrow$ **88.9%**)						
w/o X^r	56.4	1593	52.3	67.6	81.2	91.7%
CROP-PLC	**60.3**	**1634**	**56.2**	**71.8**	**83.7**	**96.5%**

Table 6: Ablation Study on the Impact of Anchor-Preserved Tokens X^r in the PLC Strategy on LLaVA-1.5-7B. Performance is compared between the full CROP-PLC and CROP-PLC(w/o X^r). The results demonstrate that when the anchor tokens X^r are discarded, the model's performance declines as it then processes visual tokens from the key region that are effectively more fragmented. The **bold values** indicate the best performance.

spatially disjointed or peripherally relevant visual tokens that offer limited effective visual guidance. In contrast, CROP's approach, by identifying and preserving contextual regions as illustrated in Figure 4, inherently leverages the contiguity of visual objects and query-specific context. This method focuses the selection on principal visual elements relevant to the query, maintaining spatial coherence and naturally filtering out irrelevant background or edge information. Consequently, CROP yields a more compact and semantically rich set of visual tokens, leading to the improved performance retention validated in our main results.

5.3 Impact of Contextual Region Preservation on Perception

To further quantify the importance of explicitly preserving contextual regional information within our PLC framework, we conducted an ablation study on LLaVA-1.5-7B. We trained a variant of the PLC architecture that omits the anchor-preserved tokens X^r and the associated fusion step (Equation 4). In this ablated setup, the visual input to the LLM comprised only the concatenated compressed representations from contextual regions $\hat{X}^{kv}$ and non-contextual regions $\hat{X}^{nkv}$.

As detailed in Table 6, removing the anchor-preserved tokens X^r led to a notable performance decrease of approximately 5% on average across benchmarks when compared to the full PLC strategy. Furthermore, this ablated PLC variant sometimes underperformed the simpler "prune" strategy (Section 5.1). This finding strongly supports our assertion that explicitly preserving a core set of coherent regional tokens, via mechanisms like the X^r fusion in PLC, is crucial for maintaining the model's perceptual capabilities, especially under aggressive token compression. This finding suggests that for VLMs, preserving the structural integrity of key visual information is as critical as retaining individual, semantically important tokens. This insight—that spatial coherence matters—is a key consideration that could inform the design of future token reduction techniques.

6 Conclusion

This work addresses the computational inefficiency in VLMs caused by excessive visual tokens. We introduced CROP, a framework that dramatically reduces token count by first identifying and then preserving contiguous, query-relevant visual regions. Our experiments demonstrate that CROP significantly outperforms existing methods. Notably, our training-free ILP strategy achieves competitive performance, offering an efficient drop-in solution for existing VLMs. Our findings underscore a critical principle for future VLM design: preserving the spatial integrity of key visual information is essential for building efficient and perceptually robust models.

Limitation

Despite the remarkable efficiency demonstrated by CROP, particularly with its training-free ILP strategy, the computational gains are primarily observed during inference. The initial fine-tuning of the PLC module, while a one-time cost, does add to the overall model development pipeline. Future work might explore methods to further streamline this initial training phase or develop entirely training-free compression strategies, both internally and externally, such as meta-learning universal compression approaches that require minimal to no task-specific fine-tuning, thereby achieving even more efficient and true "zero-shot" compression.

Acknowledgements

We would like to express our gratitude to the anonymous reviewers for their insightful feedback and constructive suggestions. This research was supported by grants from the National Natural Science Foundation of China (No. 62476275).

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei

Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Jianghao Chen, Zhenlin Wei, Zhenjiang Ren, Ziyong Li, and Jiajun Zhang. 2025a. Lr$\hat{2}$bench: Evaluating long-chain reflective reasoning capabilities of large language models via constraint satisfaction problems. *Preprint*, arXiv:2502.17848.

Jianghao Chen, Junhong Wu, Yangyifan Xu, and Jiajun Zhang. 2025b. Ladm: Long-context training data selection with attention-based dependency measurement for llms. *Preprint*, arXiv:2503.02502.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. *CoRR*, abs/2403.06764.

Wei-Ge Chen, Irina Spiridonova, Jianwei Yang, Jianfeng Gao, and Chunyuan Li. 2023. Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing. *Preprint*, arXiv:2311.00571.

Xingjian Diao, Tianzhen Yang, Chunhui Zhang, Weiyi Wu, Ming Cheng, and Jiang Gui. 2025a. Learning sparsity for effective and efficient music performance question answering. *arXiv preprint arXiv:2506.01319*.

Xingjian Diao, Chunhui Zhang, Keyi Kong, Weiyi Wu, Chiyu Ma, Zhongyu Ouyang, Peijun Qing, Soroush Vosoughi, and Jiang Gui. 2025b. Soundmind: Rl-incentivized logic reasoning for audio-language models. *arXiv preprint arXiv:2506.12935*.

Xingjian Diao, Chunhui Zhang, Tingxuan Wu, Ming Cheng, Zhongyu Ouyang, Weiyi Wu, and Jiang Gui. 2024. Learning musical representations for music performance question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Haonan Ge, Yiwei Wang, Ming-Hsuan Yang, and Yujun Cai. 2025. Mrfd: Multi-region fusion decoding with self-consistency for mitigating hallucinations in lvlms. *Preprint*, arXiv:2508.10264.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *Preprint*, arXiv:1612.00837.

Boyu Guan, Yining Zhang, Yang Zhao, and Chengqing Zong. 2025. TriFine: A large-scale dataset of vision-audio-subtitle for tri-modal machine translation and benchmark with fine-grained annotated tags. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8215–8231, Abu Dhabi, UAE. Association for Computational Linguistics.

Yuhang Han, Xuyang Liu, Zihan Zhang, Pengxiang Ding, Donglin Wang, Honggang Chen, Qingsen Yan, and Siteng Huang. 2025. Filter, correlate, compress: Training-free token reduction for mllm acceleration. *Preprint*, arXiv:2411.17686.

Wenxuan Huang, Zijie Zhai, Yunhang Shen, Shaosheng Cao, Fei Zhao, Xiangfeng Xu, Zheyu Ye, and Shaohui Lin. 2024. Dynamic-llava: Efficient multimodal large language models via dynamic vision-language context sparsification. *CoRR*, abs/2412.00876.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Preprint*, arXiv:1902.09506.

Pu Jian, Junhong Wu, Wei Sun, Chen Wang, Shuo Ren, and Jiajun Zhang. 2025a. Look again, think slowly: Enhancing visual reflection in vision-language models. *Preprint*, arXiv:2509.12132.

Pu Jian, Donglei Yu, Wen Yang, Shuo Ren, and Jiajun Zhang. 2025b. Teaching vision-language models to ask: Resolving ambiguity in visual questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3619–3638.

Pu Jian, Donglei Yu, and Jiajun Zhang. 2024. Large language models know what is key visual entity: An llm-assisted multimodal retrieval for vqa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10939–10956.

Ye Jing and Xinpei Zhao. 2024. Dq-former: Querying transformer with dynamic modality priority for cognitive-aligned multimodal emotion recognition in conversation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4795–4804.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. 2024a. Tokenpacker: Efficient visual projector for multimodal llm. *Preprint*, arXiv:2407.02392.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024b. Mini-gemini: Mining the potential of multi-modality vision language models. *Preprint*, arXiv:2403.18814.

You Li, Heyu Huang, Chi Chen, Kaiyu Huang, Chao Huang, Zonghao Guo, Zhiyuan Liu, Jinan Xu, Yuhua Li, Ruixuan Li, and Maosong Sun. 2025. Migician: Revealing the magic of free-form multi-image grounding in multimodal large language models. *Preprint*, arXiv:2501.05767.

Yupu Liang, Yaping Zhang, Cong Ma, Zhiyang Zhang, Yang Zhao, Lu Xiang, Chengqing Zong, and Yu Zhou. 2024. Document image machine translation with dynamic multi-pre-trained models assembling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7084–7095.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Ting Liu, Liangtao Shi, Richang Hong, Yue Hu, Quanjun Yin, and Linfeng Zhang. 2024b. Multi-stage vision token dropping: Towards efficient multimodal large language model. *CoRR*, abs/2411.10803.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Preprint*, arXiv:2209.09513.

Bozhi Luan, Hao Feng, Hong Chen, Yonghui Wang, Wengang Zhou, and Houqiang Li. 2024. Textcot: Zoom in for enhanced multimodal text-rich image understanding. *Preprint*, arXiv:2404.09797.

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. 2025. Argus: Vision-centric reasoning with grounded chain-of-thought. *Preprint*, arXiv:2505.23766.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. *Preprint*, arXiv:2312.02051.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Preprint*, arXiv:2403.16999.

Zhenwei Shao, Mingyang Wang, Zhou Yu, Wenwen Pan, Yan Yang, Tao Wei, Hongyuan Zhang, Ning Mao, Wei Chen, and Jun Yu. 2025. Growing a twig to accelerate large vision-language models. *Preprint*, arXiv:2503.14075.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8317–8326. Computer Vision Foundation / IEEE.

Wei Sun, Wen Yang, Pu Jian, Qianlong Du, Fuwei Cui, Shuo Ren, and Jiajun Zhang. 2025. Ktae: A model-free algorithm to key-tokens advantage estimation in mathematical reasoning. *arXiv preprint arXiv:2505.16826*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Penghao Wu and Saining Xie. 2023. V*: Guided visual search as a core mechanism in multimodal llms. *Preprint*, arXiv:2312.14135.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. 2024. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *CoRR*, abs/2410.17247.

Yangyifan Xu, Jianghao Chen, Junhong Wu, and Jiajun Zhang. 2025. Hit the sweet spot! span-level ensemble for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8314–8325.

Dawei Yan, Pengcheng Li, Yang Li, Hao Chen, Qingguo Chen, Weihua Luo, Wei Dong, Qingsen Yan, Haokui Zhang, and Chunhua Shen. 2024. Tg-llava: Text guided llava via learnable latent embeddings. *Preprint*, arXiv:2409.09564.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. 2024. Visionzip: Longer is better but not necessary in vision language models. *Preprint*, arXiv:2412.04467.

Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, and Yansong Tang. 2025. Voco-llama: Towards vision compression with large language models. *Preprint*, arXiv:2406.12275.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *ArXiv preprint*, abs/2306.13549.

Runpeng Yu, Xinyin Ma, and Xinchao Wang. 2025. Introducing visual perception token into multimodal large language model. *Preprint*, arXiv:2502.17425.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. 2025a. Mllms know where to look: Training-free perception of small visual details with multimodal llms. *Preprint*, arXiv:2502.17422.

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. 2025b. Llava-mini: Efficient image and video large multimodal models with one vision token. *CoRR*, abs/2501.03895.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A. Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. 2024. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *CoRR*, abs/2410.04417.

Yunhao Zhang, Shaonan Wang, Nan Lin, Xinyi Dong, Chong Li, and Chengqing Zong. 2025c. Discovering semantic subdimensions through disentangled conceptual representations. *Preprint*, arXiv:2508.21436.

Zhiyang Zhang, Yaping Zhang, Yupu Liang, Cong Ma, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2025d. Understand layout and translate text: Unified feature-conductive end-to-end document image translation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–18.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2025. Mlvu: Benchmarking multi-task long video understanding. *Preprint*, arXiv:2406.04264.

Appendix

A Template of the Training Data Examples

We now describe our training data format. Both the Localization Module and the PLC strategy's compression module are trained using examples where contextual regions are defined by the 8x8 grid-based block coordinates presented in Section 3.2. Table 1 provides further details on the datasets used.

Template of Training Example for Region Selection Token

User: <image> <question> Please identify the region that can help you answer the question better.
Assistant: <x_min> <y_min> <x_max> <y_max>.

B Details of Evaluation Benchmarks

In this section, we provide a brief overview of the benchmarks used in our experiments.

GQA (Hudson and Manning, 2019) serves as a benchmark for evaluating visual scene understanding and reasoning capabilities. It utilizes a combination of images, associated questions, and scene graphs, specifically testing a model's ability to grasp spatial relationships and object properties within intricate visual contexts.

MME (Fu et al., 2024) offers a comprehensive assessment of model performance across 14 distinct subtasks, which investigate both perceptual abilities and cognitive skills. The benchmark employs meticulously designed instruction-answer pairs to ensure an equitable and thorough evaluation of a model's multimodal competence. The final reported score represents the aggregate of the perception and cognition scores.

TextVQA (Singh et al., 2019) is designed to gauge a model's capacity to interpret and reason about textual elements found within images. By demanding the synthesis of visual and textual data, it stands as an important benchmark for assessing text-oriented reasoning in visual environments. For conciseness in our experimental tables, we refer to it as "VQA^{T}".

ScienceQA (Lu et al., 2022) encompasses a broad range of scientific disciplines, including natural, language, and social sciences. Its questions are structured into 26 topics, 127 categories, and 379 skills. This benchmark evaluates a model's multimodal comprehension, aptitude for multi-step reasoning, and its interpretability, thus offering a robust framework for assessing the application of scientific knowledge in visual contexts. In our experiments, we focus exclusively on the image-based samples, denoted as "SQA^{I}" in the experimental tables.

VQA-v2 (Goyal et al., 2017) is a comprehensive benchmark consisting of 265,000 images that capture real-world scenes and objects. Each image is presented with open-ended questions, and for each question, ten ground truth answers provided by humans are available.

POPE (Yue et al., 2024) is focused on evaluating object hallucination by presenting binary questions concerning the existence of objects in images. It utilizes metrics such as Accuracy, Recall, Precision, and F1 score, applied over three different sampling methodologies. The score reported reflects the mean accuracy across these three methods: adversarial, random, and popular.

C Benchmarks and Metrics for Evaluation

The benchmarks utilized in our study, along with their respective evaluation metrics, adhere to the official guidelines provided by each benchmark and the official evaluation scripts from the LLaVA model. For the VQA datasets, we assessed the zero-shot question-answering performance using a single input image. All results reported in this paper are averaged across multiple experimental runs.

D More Experimental Results

D.1 Results on Qwen

After demonstrating the effectiveness of the CROP framework on the LLaVA family of models, we further evaluate its performance on the Qwen series. The Qwen2-VL model adaptively determines the number of visual tokens based on the resolution of the input image, allowing it to process high-resolution inputs and exhibit strong visual understanding capabilities. As shown in Table 7, we evaluate the performance of CROP-ILP under four different token retention settings. Even with pruning rates ranging from 50% to 77.8%, Qwen2-VL-7B retains between 98.1% and 95.8% of the original performance. These results demonstrate the effi-

Method	GQA	MME	VQA^T	SQA	POPE	RelAcc
Dynamic Number of Tokens (**100%**)						
Qwen2-VL	61.58	2269	83.92	84.34	88.35	100%
Retain Averaged 50% Tokens						
FastV	55.50	**2251**	79.90	78.90	82.99	94.4%
CROP-ILP	**61.37**	2143	**81.80**	**83.42**	**88.41**	**98.1%**
Retain Averaged 40% Tokens						
FastV	55.58	**2216**	76.52	77.46	81.16	92.6%
CROP-ILP	**60.67**	2066	**81.61**	**83.07**	**88.23**	**97.0%**
Retain Averaged 33.3% Tokens						
FastV	54.64	**2174**	73.12	75.13	80.36	90.3%
CROP-ILP	**60.21**	2050	**80.61**	**82.62**	**88.06**	**96.4%**
Retain Averaged 22.2% Tokens						
FastV	53.30	**2041**	71.56	72.84	76.15	86.9%
CROP-ILP	**59.85**	2013	**80.34**	**82.60**	**87.86**	**95.8%**

Table 7: Performance of CROP-ILP with Qwen2-VL-7B on Various VLM Benchmarks. The **bold values** indicate the best performance.

Method	GQA	MME	VQA^T	SQA	POPE	RelAcc
Upper Bound, 576 Tokens (**100%**)						
LLaVA-1.5	61.94	1862	58.20	69.54	85.93	100%
Retain Averaged 64 Tokens (↓ **88.9%**)						
CROP-ILP(K=32)	61.96	1866	58.18	70.95	85.93	100.5%
CROP-ILP(K=31)	61.94	1869	57.92	70.95	85.99	100.4%
CROP-ILP(K=30)	61.99	1855	58.04	70.97	86.01	100.3%
CROP-ILP(K=25)	61.97	1854	57.77	70.97	86.09	100.2%
CROP-ILP(K=20)	61.63	1871	57.48	70.93	86.09	100.2%
CROP-ILP(K=15)	61.35	1869	56.60	71.00	85.84	99.7%
CROP-ILP(K=10)	60.11	1749	55.78	70.88	83.68	97.2%
CROP-ILP(K=7)	59.92	1775	54.80	71.47	84.28	97.4%
CROP-ILP(K=5)	59.62	1713	54.59	71.47	83.71	96.5%
CROP-ILP(K=3)	59.58	1670	54.46	71.23	83.69	95.9%
CROP-ILP(K=2)	59.59	1675	54.92	71.54	83.57	96.1%
CROP-ILP(K=1)	59.62	1665	54.24	70.88	83.55	95.6%
prune	55.03	1625	54.13	69.92	83.84	93.5%

Table 8: Ablation study on the pruning layer K in the LLaVA-1.5-7B model. The results demonstrate that applying the CROP-ILP method for pruning across different layers of the model largely preserves its original performance.

Sample Count	Mean Recall	Recall >0.5	Recall >0.7	Recall >0.9
3000	0.9477	2871	2802	2706

Table 9: The Area-based Recall value of the contextual region localized by the Localization Module.

ciency and robustness of our method on the Qwen architecture, and further highlight the strong generalization ability of the proposed CROP framework across diverse multimodal model families.

D.2 Ablation Study on the Pruning Layer

In the aforementioned CROP-ILP experiments, we set the pruning layer K=2 to ensure a fair comparison with other methods that prune visual tokens in the early layers of the model. To further examine the impact of pruning at different layers, we conducted an ablation study on LLaVA-1.5-7B using the ILP method, where the visual tokens were pruned to an average of 64. As shown in the Table 8, increasing K allows the VLM to retain more visual information, resulting in a general improvement in overall performance, and in fact, pruning at the top layers can even lead to performance gains. On the other hand, pruning in the early layers yields substantial efficiency improvements but inevitably introduces minor performance drops. In practical scenarios, one may balance these trade-offs and select an appropriate pruning layer K according to the application requirements.

D.3 Robustness of the Localization Module

To validate the robustness of our Localization Module, we randomly selected 3000 samples from TextVQA set and manually annotated the ground truth regions. In the following, GT denotes the manually labeled ground truth region, and $Pred$ denotes the region predicted by our Localization Module. We used Area-based Recall to quantify localization quality, which is defined as the ratio of the intersection area between GT and $Pred$ to the area of GT. This metric reflects how accurately the predicted region capture the true relevant visual context. As shown in the Table 9, the results show that our Localization Module achieves high perception accuracy, with strong overlap between the predicted and human-annotated contextual regions.

To measure how localization quality affects model performance, we selected two alternative regions in each image: a center region($Center$) and a randomly chosen region($Random$), both resized to the same size as the predicted region. We then applied ILP pruning with pruning layer K = 2 on several VLMs. The results are shown in the Table 10.

They demonstrate that localization quality has a substantial impact on pruning performance. The predicted regions from our Localization Module closely match the performance of GT regions, whereas the $Center$ and $Random$ regions cause significant drops in model performance.

LLaVA-1.5-7B baseline	*GT*	*Pred*	*Center*	*Random*
52.32	53.20	53.14	47.35	45.76
LLaVA-1.5-13B baseline	***GT***	***Pred***	***Center***	***Random***
55.90	55.85	55.81	49.78	48.47
Qwen2VL-7B baseline	***GT***	***Pred***	***Center***	***Random***
83.75	81.93	81.72	66.44	62.67

Table 10: The performance of the CROP-ILP strategy on LLaVA and Qwen models with different region selection strategies.

E More Visualized Results

In this section, we present several visual examples of the CROP method to provide a more intuitive understanding of the importance of contextual regions in visual token pruning. All experiments are conducted using LLaVA-1.5-7B as the base Visual Language Model. We have gathered a collection of both successful and unsuccessful cases, including those involving routine questions, questions involving object relationships, and more global questions. Overall, in the majority of cases, CROP and the LLaVA model demonstrated consistent performance. In tasks where contextual regions are crucial, we found that by helping the VLM locate the key information, it was able to correctly answer questions that would otherwise be incorrectly answered, even after pruning most of the visual tokens. This is because we effectively eliminated unnecessary visual distractions, allowing the VLM to focus on the key visual entities, as shown in Figure 5.

However, for tasks involving object relationships, the VLM requires awareness of multiple objects or parts of them to provide correct answers. If too much visual information is discarded, it impairs the VLM's judgment. We present both successful and unsuccessful cases in this regard. In the three examples at the bottom of Figure 5, the questions involve two objects. Our Localization Module preserved the most critical information in the contextual region, retaining only part of the objects involved in the relationships. Despite this, the VLM was still able to make the correct judgment and answer accurately. In contrast, in the example shown in Figure 6, the VLM lost too much object information, preventing it from making an accurate judgment and resulting in an incorrect answer.

Furthermore, for tasks requiring global information, the VLM is likely to provide an incorrect answer if the contextual region does not encompass the necessary global context, as shown in Figure 7. To address these types of problems, we will explore the use of multiple contextual regions in future work, which will enhance the granularity of localization and help the VLM retain key visual objects as well as global visual information.

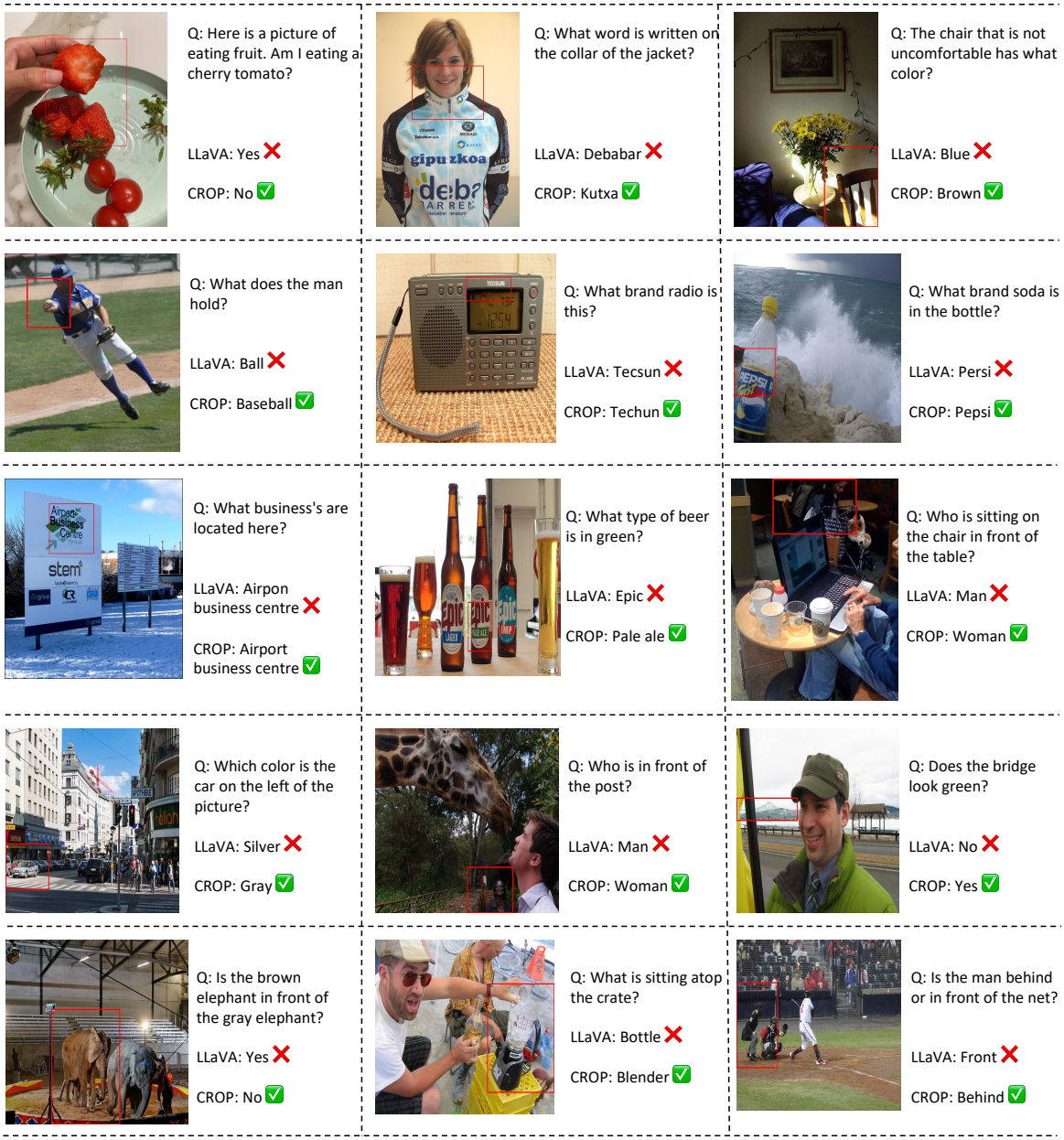

Figure 5: Examples of the CROP-ILP strategy applied to the LLaVA-1.5-7B model, including some routine questions and questions involving object relationships. The responses of both the base LLaVA model and the LLaVA model using CROP are presented, with check marks and cross marks indicating whether each response is correct or incorrect.

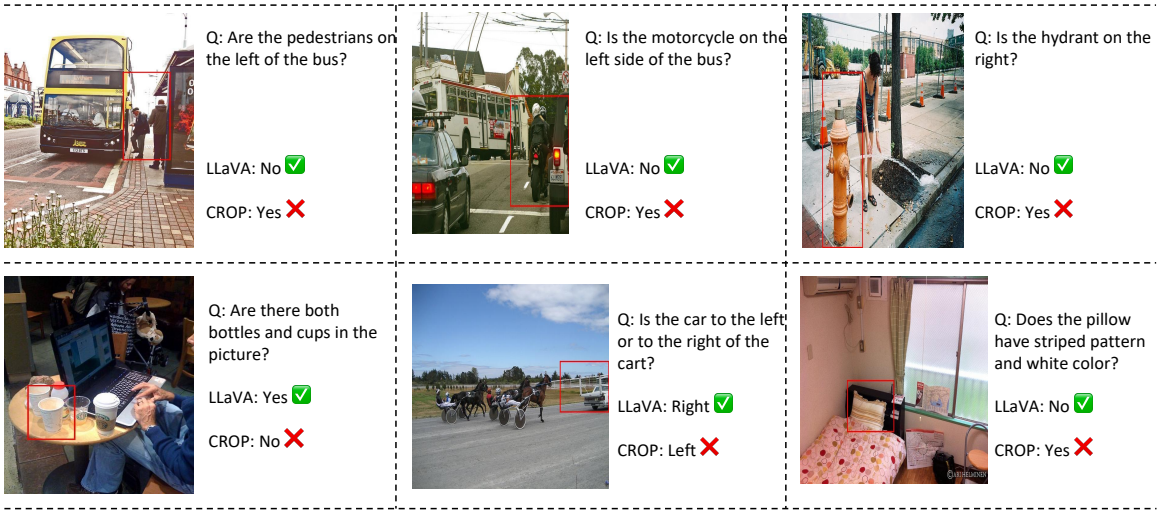

Figure 6: Examples of the CROP-ILP strategy applied to the LLaVA-1.5-7B model, including some questions involving object relationships. The responses of both the base LLaVA model and the LLaVA model using CROP are presented, with check marks and cross marks indicating whether each response is correct or incorrect.

Figure 7: Examples of the CROP-ILP strategy applied to the LLaVA-1.5-7B model, including some questions requiring global information. The responses of both the base LLaVA model and the LLaVA model using CROP are presented, with check marks and cross marks indicating whether each response is correct or incorrect.